National Statistical Service

Confidentiality
Information Sheet

Glossary

Administrative data

Information (including personal information) collected by agencies for the administration of programs, policies or services and with the potential to be used for statistical purposes. Administrative data is one type of **microdata.**

Aggregate data

Aggregate data are produced by grouping information into categories and aggregating values within these categories. For example, a count of the number of people of a particular age (obtained from the question 'In what year were you born?').

Aggregate data is typically presented in tables. Aggregate data is also referred to as **tabular data** or **macrodata.**

Anonymised data

This term is most commonly used to refer to data from which direct identifiers have been removed (de-identified data), but it is sometimes also used to refer to confidentialised data. To avoid confusion the more specific terms **de-identified data** and **confidentialised data** are used in the Confidentiality Information Series.

Cell concentration rule

See **Cell dominance rule**.

Cell dominance rule

A rule commonly applied to cells in a table to assess whether a cell may enable identification. The cell dominance rule (also called the cell concentration rule) is used to identify cells where a small number of data providers contribute a large percentage to the cell. If a cell fails this rule further investigation or action is needed to ensure that identification is unlikely.

For more information see *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

Confidential data

Data that will *allow identification* of an individual or organisation, either directly or indirectly.

Confidentialise

To remove or alter information, or collapse detail within a dataset, to ensure that no person or organisation is likely to be identified in the data (directly or indirectly).

For more information see *Confidentiality Information Sheet 1: 'Confidentiality: what is it and why is it important?'* and *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles.'*

Confidentialised Unit Record Files (CURFs)

Confidentialised Unit Record Files (CURFs) are files containing microdata that have been de-identified and modified to protect individuals or organisations from either direct or indirect identification.

Confidentiality

'An obligation to the provider of information to maintain the secrecy of that information.'

Source: UN Economic Commission for Europe, 2009.

For more information see *Confidentiality Information Sheet 1: 'Confidentiality: what is it and why is it important?'*

Confidentiality rules

Rules that are applied to each cell in a table to identify table cells that pose a risk of identification (disclosure). Two common rules are the **frequency rule** and the **cell dominance rule.**

Data custodian

The organisation or agency which is responsible for the collection, use and disclosure of information in a dataset. Data custodians have an obligation to keep the confidential information they are entrusted with secret.

Data laboratories

On-site data laboratories provide access to detailed microdata at a secure site controlled by the data custodian.

Data modification

See **Perturbation.**

Data provider

An individual, household, business or other organisation which supplies data either for statistical or administrative purposes.

Data reduction

A technique used to confidentialise data. Data reduction methods aim to control or limit the amount of detail available to avoid identification of a particular individual or business. Data reduction methods include combining categories of information or suppressing information for unsafe cells.

For more information about techniques to confidentialise data see: *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

Data rounding

Involves slightly altering small cells in a table to ensure results from analysis based on the data are not significantly affected, but the original values cannot be known with certainty. Data rounding may be random or controlled.

For more information about data rounding techniques see: *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

De-identified data

Data that have had any identifiers removed. May also be referred to as **unidentified data**.

Direct identification

Occurs when a direct **identifier** is included with the data that can be used to establish the identity of an individual or organisation.

Disclosure

Disclosure occurs when a person or an organisation recognises or learns something that they did not already know about another person or organisation through released data.

Disclosure control

Managing the risks of an individual or organisation being identified either directly or indirectly.

Frequency rule

A rule commonly applied to cells in a table to assess whether a cell may enable identification. The **frequency rule** (also called the **threshold rule**) sets a threshold value for the minimum number of individuals or businesses in any cell. Common threshold values are 3, 5 and 10. If a cell fails this rule further investigation or action is needed to ensure that identification is unlikely.

Identifiable

Able to be identified either directly or indirectly.

Identification

See **Direct identification, Identified data** and **Indirect identification.**

Glossary

Identified data

Data that include an identifier.

Identifier(s)

An identifier (direct identifier) is information that directly establishes the identity of an individual or organisation. Examples of identifiers are: name, address, driver's licence number, Medicare number and Australian Business Number.

Indirect identification

Occurs when the identity of an individual or organisation is disclosed, not through the use of direct identifiers, but through a combination of unique characteristics.

Macrodata

See **Aggregate data.**

Microdata

Unit record data where each record represents observations for an individual or organisation. Unit record data may contain individual responses to questions on a survey questionnaire or administrative forms. For example, answers given to the question 'In what year were you born?'.

Outlier

An unusual value that is correctly reported but is not typical of the rest of the population.

Personal information

'Information or an opinion (including information or an opinion forming part of a database), whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.' *(Privacy Act 1988)*

Personal information is information that identifies, or could identify a person. There are some obvious examples of personal information, such as name or address. Personal information can also include medical records, bank account details, photos, videos, and even information about what a person likes, their opinions and where they work – basically, any information through which a person is reasonably identifiable.

Information does not have to include a name to be personal information. For example, in some cases, date of birth and post code may be enough to identify someone.

Perturbation

A technique used to confidentialise data. Perturbation is a data modification method that involves changing the data slightly to reduce the risk of disclosure while retaining as much content and structure as possible.

Perturbation techniques include data rounding or data swapping.

For further information about techniques to confidentialise data see: *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

Privacy

The individual's right to have their personal information managed so that it is kept confidential except where informed consent has been given to release the information, or a legal authority exists, in accordance with the requirements of the *Privacy Act 1988*.

Provider

An individual, household, business or other organisation which provides data either to statistical collections or administrative collections. May also be referred to as a **respondent.**

Record attack

Occurs when a user tries to find a particular person or organisation with a set of characteristics known to the user.

Remote access facility

Remote access facilities are used by agencies around the world to enable approved researchers to submit data queries through a secure internet-based interface from their desktop. The request is run against a confidentialised unit record file, which is securely stored within the data custodian's computing environment.

Glossary

Remarkable characteristics

The presence of a rare characteristic in the data can pose an identification risk, depending on how **remarkable** (or extraordinary or noticeable) the characteristic is. This might include people in unusual jobs, very large families or young people with very high educational qualifications.

Risk management

In the context of confidentiality, risk management involves identification and management of the risk of disclosure in accordance with the impact and likelihood of a disclosure occurring and within the constraints provided by legislation and policies.

Further information about the process of managing microdata disclosure risks is in *Confidentiality Information Sheet 5 – 'Confidentiality: managing the risk of disclosure in the release of microdata.'*

Security

Safe storage and access to held data, including physical security of buildings and IT security.

Spontaneous recognition

An identification made without any deliberate attempt.

Statistical purposes

Purposes which support the collection, storage, compilation, analysis and transformation of data for the production of statistical output, and the dissemination of those outputs and the information describing them.

This means that information cannot be used for administrative, regulatory, law enforcement or other purpose that affects the rights, privileges or benefits of particular individuals or organisations.

Suppression

Data suppression involves not releasing information that is considered unsafe because it fails confidentiality rules being applied.

For further information about techniques to suppress data see: *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

Tabular data

See **Aggregate data**.

Threshold rule

See **Frequency rule**.

Unidentified data

See **De-identified data**.

Uniqueness

Used to characterise the situation where an individual can be distinguished from all other members in a population or sample in terms of information available on **microdata** records. The existence of uniqueness is determined by the size of the population or sample, the degree to which it is segmented by geographic information and the number and detail of characteristics provided for each unit in the dataset.

Unit record data

See **Microdata**.